Notes on the birth-death prior with fossil calibrations for Bayesian estimation of species divergence times

Mario dos Reis[1,2]

26th April 2016

1. School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK.
2. Department of Genetics, Evolution and Environment, University College London, Gower Street, London, W1C 6BT.
Email: m.dosreisbarros@qmul.ac.uk
Phone: +44-20-7882-6108

Keywords: phylogeny, divergence time, molecular clock, time prior, Bayesian method, birth-death process

Abstract

Constructing a multi-dimensional prior on the times of divergence (the node ages) of species in a phylogeny is not a trivial task, in particular if the prior density is the result of combining different sources of information such as a speciation process with fossil calibration densities. Yang and Rannala (2006, Mol. Biol. Evol., 23: 212–226) laid out the general approach to combine the birth-death process with arbitrary fossil-based densities to construct a prior on divergence times. They achieved this by calculating the density of node ages without calibrations conditioned on the ages of the calibrated nodes. Here I show that the conditional density obtained by Yang and Rannala is misspecified. The misspecified density can sometimes be quite strange-looking and can lead to unintentionally informative priors on node ages without fossil calibrations. I derive the correct density and provide a few illustrative examples. Calculation of the density involves a sum over a large set of labelled histories, and so obtaining the density in a computer program seems hard at the moment. A general algorithm that may provide a way forward is given.

1 Introduction

There has been much interest in using the Bayesian method to estimate the times of divergence (the node ages) of species in phylogenies [1]. However, in a Bayesian analysis, specifying the prior distribution on the ages of nodes is not a trivial task: for a phylogeny with s species, there are $s-1$ node ages to be estimated, and thus, a probability distribution with $s-1$ dimensions needs to be constructed. Such high dimensional priors may be hard to specify, in particular if the prior is built from different sources of information, such as when combining information from the fossil record with a speciation process (such as the birth-death process [2]).

For example, in the first Bayesian method of molecular clock dating, Thorne et al. [3] used the Yule process (a speciation process) to construct the prior on the node ages, but no attempt was made to combine the process with node age calibrations based on the fossil record. Later, Kishino et al. [4] used a gamma density to specify the prior on the age of the phylogeny's root, and then used the Dirichlet distribution to construct the prior on the remaining node ages. They innovatively used the ages of fossils as minimum or maximum constraints on the ages of nodes (that is, by truncating the Dirichlet distribution). Thus, in Kishino et al.'s method, information from the fossil record can be used to inform the time prior throughout the phylogeny, but arbitrary fossil-based calibration densities on node ages could not be used.

Bayesian methods to construct the prior on node ages by combining a speciation process with arbitrary fossil-based distributions were developed later. For example, Drummond et al. [5] introduced one such method, although no mathematical details on how their prior is constructed were initially given. Details of prior construction were given later [6]: the prior density is constructed by simply multiplying the speciation process density with the arbitrary probability densities on nodes with fossil calibrations. This approach to construct the prior seems undesirable as it "does not follow the rules of probability calculus" [6].

Yang and Rannala [2] laid out the correct approach to combine a speciation process density with arbitrary probability densities to construct the time prior. Let t_1 be the age of

the root, and let $\mathbf{t}_{\bar{1}} = (t_2, \ldots, t_{s-1})$ be the ages of the remaining $s-2$ nodes in a phylogeny of s species. Write $f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1)$ for the birth-death density of node ages conditioned on the age of the root [2]. Let $f(t_1)$ be a probability density (such as the gamma) describing the age of the root. A joint prior of divergence times can thus be constructed as

$$f(t_1, \ldots, t_{s-1}) = f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1) \times f(t_1).$$

Now suppose fossil calibration information is available for a subset $\mathbf{t}_{\mathrm{c}}$ of the node ages (other than the root), while $\mathbf{t}_{\bar{\mathrm{c}}}$ are the node ages without fossil calibrations, such that $\mathbf{t}_{\bar{1}} = (\mathbf{t}_{\mathrm{c}}, \mathbf{t}_{\bar{\mathrm{c}}})$. Write $f(\mathbf{t}_{\mathrm{c}})$ for the joint probability density that summarises the fossil information about $\mathbf{t}_{\mathrm{c}}$. A joint time prior on all node ages that combines the birth-death process density and the fossil-based densities can be constructed as

$$\begin{aligned} f(t_1, \ldots, t_{s-1}) &= \frac{f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}}, \mathbf{t}_{\mathrm{c}} \mid t_1)}{f_{\mathrm{BD}}(\mathbf{t}_{\mathrm{c}} \mid t_1)} \times f(\mathbf{t}_{\mathrm{c}}) \times f(t_1) \\ &= f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}} \mid \mathbf{t}_{\mathrm{c}}, t_1) \times f(\mathbf{t}_{\mathrm{c}}) \times f(t_1). \end{aligned} \quad (1)$$

Thus, to calculate the time prior using fossil calibrations and the birth-death process, one must calculate: (i) $f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}}, \mathbf{t}_{\mathrm{c}} \mid t_1)$, (ii) $f_{\mathrm{BD}}(\mathbf{t}_{\mathrm{c}} \mid t_1)$, and (iii) $f(\mathbf{t}_{\mathrm{c}}) \times f(t_1)$. Calculating (i) and (iii) is straightforward, however, it turns out that calculating (ii) is in general very hard. In fact, here I show that the procedure proposed by Yang and Rannala to obtain (ii) leads to a misspecified density. Thus Yang and Rannala's resulting conditional density for node ages without calibrations, $f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}} \mid \mathbf{t}_{\mathrm{c}}, t_1)$, is also misspecified. The problem is moderate, in the sense that the fossil calibration densities are not affected, and estimated divergence times on the phylogeny will be appropriately constrained by the fossil calibrations specified by the user. However, the misspecified $f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}} \mid \mathbf{t}_{\mathrm{c}}, t_1)$ density can look quite strange and sometimes can have multiple modes, and may thus lead to a multi-modal posterior distribution for the ages of nodes without fossil calibrations. Such misspecified densities may be biologically unrealistic. An example of the misspecified prior is given in figure 1.

The prior of Eq. (1), using the misspecified densities, has been implemented in the computer program MCMCTree [7] for molecular clock dating of phylogenies. The program is popular and has been used, for example, to estimate node ages in phylogenies using genome-scale alignments (e.g. [8, 9, 10]). Given the popularity of MCMCTree, and given the limitations of other methods (e.g., [3, 4, 5]), it is important to obtain the correct prior density of divergence times with fossil calibrations.

Thus the purpose of this paper is to derive, from first principles, the prior density of divergence times under the birth-death process with fossil calibrations of Eq. (1). Here I show that calculation of the correct density in small phylogenies is straightforward, and I give a general form of the density for arbitrarily large phylogenies. Unfortunately, the density involves a sum over the set of possible labelled histories in the phylogeny (i.e. the possible orderings of node ages given the rooted tree). This set can be explosively large in phylogenies of many species, and thus application of the general form of the density is impractical. Thus I also sketch out a more efficient tree traversal algorithm that can be used to calculate the density. A few special cases where the MCMCTree program calculates the density correctly are pointed out, and brief recommendations to construct a reasonable prior in other cases are given.

The procedure of Yang and Rannala [2] is used to construct the time prior in phylogenies of extant (or contemporaneous) species. In some cases, data for extint species may be available (for example, molecular data for viruses sampled through time [11] or morphological data for fossil species [12]), or data about the temporal sampling frequency of fossils may also be available (e.g. [13, 14]). In such cases it is not necessary to use fossil calibrations densities. The birth-death process conditioned on the extant and extinct species can then be used to construct the time prior ([11, 14]). Development of the time prior in the variety of cases that may arise is an exciting and fast-paced area of research. The focus of this paper is on the special case of phylogenies of extant species with fossil calibration densities [2].

2 The birth-death process with species sampling

The birth-death process with species sampling was introduced by Yang and Rannala [15] for Bayesian estimation of phylogenies. Consider estimating the tree topology and divergence times for s contemporaneous species. The *ordered* divergence times are $t_{(1)} < t_{(2)} < \cdots < t_{(s-1)}$, where $t_{(s-1)}$ is the age of the root. Under the birth-death process, the joint (prior) distribution of a labelled history π (an ordering of node ages for topology τ) and a set of ordered times conditioned on the age of the root and the s species is

$$\begin{aligned} f_{\mathrm{BD}}(\pi, t_{(1)}, \ldots, t_{(s-2)} \mid t_{(s-1)}, s) &= \Pr(\pi) \times f_{\mathrm{BD}}(t_{(1)}, \ldots, t_{(s-2)} \mid t_{(s-1)}), \\ &= \frac{1}{N_H} \times (s-2)! \prod_{i=1}^{s-2} g(t_{(i)} \mid t_{(s-1)}), \end{aligned} \tag{2}$$

(Eq. 5 in [15]), where $g(t_{(i)} \mid t_{(s-1)})$ is the birth-death kernel density (Eq. 4 in [2]), and $N_H = s!(s-1)!/2^{s-1}$ is the total number of labelled histories given s species. In the birth-death process all histories have the same probability of being sampled, and thus $\Pr(\pi) = 1/N_H$.

For example, consider the case of four species. There are $N_H = 4!3!/2^3 = 18$ labelled histories, corresponding to 15 distinct rooted tree topologies. Figure 2 lists two of the histories for one topology. Now consider the smallest time in the phylogeny, $t_{(1)}$, and the age of the last common ancestor of species a and b, $t_{a,b}$ (we call $t_{a,b}$ a *labelled* time). In some labelled histories, $t_{(1)} = t_{a,b}$, but in general this is not the case (figure 2). Clearly, $T_{(1)}$ and $T_{a,b}$ are different random variables. Thus, in order to adapt the birth-death process to estimation of divergence times on fixed topologies with fossil calibrations, we need to derive the density of the labelled times, so that we can apply fossil calibrations on them.

For example, consider calculation of the distribution of $t_{a,b}$, given topology τ (figure 2), $f_{\mathrm{BD}}(t_{a,b} \mid \tau)$. Half of the time, we will sample history π_1 (i.e. $\Pr(\pi_1 \mid \tau) = 1/2$), where $t_{a,b} < t_{c,d}$, and thus $t_{a,b}$ will have the density of an order statistic of rank 1 (the minimum, [16]), and so $f_{\mathrm{BD}}(t_{a,b} \mid \pi_1) = 2g(t_{a,b})[1 - G(t_{a,b})]$, while the other half of the time, we will sample π_2 and thus $t_{a,b}$ will have the density of the maximum, $f_{\mathrm{BD}}(t_{a,b} \mid \pi_2) = 2g(t_{a,b})G(t_{a,b})$. Thus the density is

$$\begin{aligned} f_{\mathrm{BD}}(t_{a,b} \mid \tau) &= \Pr(\pi_1 \mid \tau) f_{\mathrm{BD}}(t_{a,b} \mid \pi_1) + \\ &\quad \Pr(\pi_2 \mid \tau) f_{\mathrm{BD}}(t_{a,b} \mid \pi_2) \\ &= g(t_{a,b}). \end{aligned}$$

To obtain the joint density of $t_{a,b}$ and $t_{c,d}$ given τ, note that there are 2! ways to order the two times, but "only" 2 of those (the two labeled histories) are compatible with the tree, this leads to

$$f_{\mathrm{BD}}(t_{a,b}, t_{c,d} \mid \tau_1) = 2!/2 \times g(t_{a,b}) g(t_{c,d}).$$

This observation can be generalised to any tree of s species to obtain the joint density for the ages of the $s-2$ internal nodes conditioned on the age of the root. One can then use first principles of probability theory to derive the marginal densities of the sets of nodes with fossil calibrations. This is the topic of the next section.

3 Birth-death prior of times with fossil calibrations

Consider a phylogeny of s species, where the topology, τ, of the phylogeny is known. The $s-1$ labelled divergence times are $\mathbf{t} = (t_1, \ldots, t_{s-1})$, where t_1 is the age of the root. The joint prior of $\mathbf{t}_{\bar{1}} = (t_2, \ldots, t_{s-1})$ conditioned on the age of the root t_1 and the tree topology τ is

$$f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1) = K \prod_{i=2}^{s-1} g(t_i \mid t_1) \times \mathbb{I}_\tau, \tag{3}$$

where the indicator function $\mathbb{I}_\tau = 0$ if any node is older than its parent, and $\mathbb{I}_\tau = 1$ if otherwise. K is the normalising constant:

$$K = 1 / \int_{\mathbf{T}_{\bar{1}}} \prod_{i=2}^{s-1} g(t_i \mid t_1) \times \mathbb{I}_\tau \, \mathrm{d}\mathbf{t}_{\bar{1}} = (s-2)!/n_H. \tag{4}$$

where n_{H} is the number of labelled histories given the rooted tree. To derive K note that there are $(s-2)!$ ways to order the $s-2$ internal node ages, but only n_H of those are compatible with the tree topology, thus resulting in Eq. (4).

The joint time prior for all the node ages given the tree topology is then

$$f(\mathbf{t}) = f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1) f(t_1),$$

where $f(t_1)$ is the fossil-based calibration density of the root age.

Now suppose that, apart from the root, there is an additional set of nodes, $\mathbf{t}_{\mathrm{c}}$, with calibration information. The set of nodes with no calibrations is $\mathbf{t}_{\bar{\mathrm{c}}}$, so that $\mathbf{t}_{\bar{1}} = (\mathbf{t}_{\bar{\mathrm{c}}}, \mathbf{t}_{\mathrm{c}})$. The calibrated nodes have joint calibration density $f(\mathbf{t}_{\mathrm{c}})$. The time prior is then defined as

$$f(\mathbf{t}) = f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}} \mid \mathbf{t}_{\mathrm{c}}, t_1) \times f(\mathbf{t}_{\mathrm{c}}) \times f(t_1),$$

where

$$f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}} \mid \mathbf{t}_{\mathrm{c}}, t_1) = \frac{f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}}, \mathbf{t}_{\mathrm{c}} \mid t_1)}{f_{\mathrm{BD}}(\mathbf{t}_{\mathrm{c}} \mid t_1)}, \tag{5}$$

and

$$f_{\mathrm{BD}}(\mathbf{t}_{\mathrm{c}} \mid t_1) = \int_{\mathbf{T}_{\bar{\mathrm{C}}}} f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1) \, \mathrm{d}\mathbf{t}_{\bar{\mathrm{c}}}. \tag{6}$$

We can obtain a general expression for Eq. (6) by using the theory of order statistics. Consider a particular labelled history π given the rooted tree. Write $t_{(i)}$ for the node age

with rank i. For example, for history $\pi = (t_2 < t_3)$ in the tree of figure 1, the rank of t_2 is 1 (because t_2 is the youngest node), and so $t_{(1)} = t_2$ and $t_{(2)} = t_3$. The brackets around the subscripts are used to emphasise that the variables are ordered. Let k be the number of nodes with fossil calibrations, and let $t_{(c_1)} < \cdots < t_{(c_k)}$, be the ranked calibrated times. The ordered node ages form a set of ordered statistics given history π. Thus, the joint density of $\mathbf{t}_c$ conditioned on history π, is the joint density of the subset of order statistics, $t_{(c_1)}, \ldots, t_{(c_k)}$. This density is given (after suitably defining $t_{(c_0)} = 0, t_{(c_{k+1})} = t_1, c_0 = 0, c_{k+1} = s-1$) by

$$\begin{aligned} f_{\mathrm{BD}}(\mathbf{t}_c \mid t_1, \pi) &= \int_{\mathbf{T}_{\bar{c}}} f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1, \pi)\, \mathrm{d}\mathbf{t}_{\bar{c}} \\ &= (s-2)! \times \prod_{j=1}^{k} g(t_{(c_j)} \mid t_1) \times \\ &\quad \prod_{j=0}^{k} \left\{ \frac{[G(t_{(c_j+1)} \mid t_1) - G(t_{(c_j)} \mid t_1)]^{c_{j+1}-c_j-1}}{(c_{j+1}-c_j-1)!} \right\} \end{aligned} \tag{7}$$

(p. 12 in [16]), where $G(t_i \mid t_1)$ is the distribution function of the birth-death kernel density, $g(t_i \mid t_1)$. Now note that the integration volume of Eq. (6) can be "sliced" such that each slice corresponds to a particular labelled history π. Therefore, the integral of Eq (6) can be expressed as a sum of integrals over labelled histories, i.e. over the slices (Eq. 7). This gives

$$\begin{aligned} f_{\mathrm{BD}}(\mathbf{t}_c \mid t_1) &= \int_{\mathbf{T}_{\bar{c}}} f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1)\, \mathrm{d}\mathbf{t}_{\bar{c}} \\ &= \sum_{\pi} \int_{\mathbf{T}_{\bar{c}}} f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1, \pi)\, \mathrm{d}\mathbf{t}_{\bar{c}} \\ &= \sum_{\pi} f_{\mathrm{BD}}(\mathbf{t}_c \mid t_1, \pi). \end{aligned} \tag{8}$$

Calculation of the conditional density $f_{\mathrm{BD}}(\mathbf{t}_c \mid t_1)$ by using Eq. (8) is, in general, not very practical. In trees of many species, the number of labelled histories may be so explosively large that the sum may not be computed. An algorithm to calculate Eq. (6) on a tree is given later.

4 Misspecification of the time prior in MCMCTree

As mentioned, the ordered times, $t_{(i)}$, and the labelled times, t_j, are different random variables, and so it would be inappropriate to assume that they have the same probability distribution. Yet, this is what Yang and Rannala [2] assumed when they adapted Eq. (2) to estimation of divergence times on a fixed topology. They equated the joint density of the set of labelled times (i.e. times that refer to a specific common ancestor, such as $t_{a,b}$) on a fixed topology, with the joint density of ordered times on a labelled history:

$$\begin{aligned} f^{*}_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1, \tau) &= f_{\mathrm{BD}}(t_{(1)}, \ldots, t_{(s-2)} \mid t_{(s-1)}, \pi) \text{ if } \pi, \\ &= (s-1)! \prod_{i=2}^{s-1} g(t_i \mid t_1), \end{aligned} \tag{9}$$

thus the normalising constant, $K^* = (s-2)!$, is incorrect (Eq. 9 in [2], cf. Eq. 3 here). They then obtained a misspecified marginal density for calibrated times by integrating the density over the ordered times:

$$\begin{aligned} f^*_{\mathrm{BD}}(\mathbf{t}_{\mathrm{c}} \mid t_1, \tau) &= \int_{\mathbf{T}_{\bar{\mathrm{c}}}} f_{\mathrm{BD}}(t_{(1)}, \ldots, t_{(s-2)} \mid t_{(s-1)}, \pi)\, \mathrm{d}\mathbf{t}_{\bar{\mathrm{c}}} \text{ if } \pi, \\ &= f_{\mathrm{BD}}(\mathbf{t}_{\mathrm{c}} \mid t_1, \pi) \end{aligned} \tag{10}$$

(Eq. 11 in [2], cf. Eq. 7 and 8 here). The asterisk is used to indicate that the densities are misspecified.

The misspecified prior is implemented in the computer program MCMCTree [7], which performs Bayesian estimation of divergence times on phylogenies by MCMC sampling. Misspecification of the integration constant, K, is unimportant, given that this constant cancels out during MCMC sampling on fixed topologies. However, the use of the misspecified conditional density is a more serious problem, as it can lead to strange time priors that do not reflect the true density under the birth-death process. Note that MCMCTree does not need to calculate $f_{\mathrm{BD}}(\mathbf{t}_{\mathrm{c}} \mid t_1, \pi)$ for all labelled histories π, it only does so for the particular labelled history being proposed during MCMC sampling. Thus, MCMCTree can perform much faster MCMC sampling using the misspecified density of Eq. (10) than how it would perform if it sampled the correct distribution (Eq. 8).

For some special cases, the misspecified (Eqs. 9,10) and correct (Eqs. 3,6) densities give the same result. Thus, for such cases, MCMCTree (version 4.8 at the time of writing) is guaranteed to calculate the birth-death prior with fossil calibrations correctly. The important cases are:

1. When there is a single calibration on the age of the root, and the birth-death process is used to specify the prior on all remaining nodes. In this case only K is miscalculated, but as mentioned, this is unimportant.

2. For comb phylogenies, irrespective of the configuration of the fossil calibrations. In a comb phylogeny, each node has only one other internal node as its child (the other child is a tip), and therefore there is a single labelled history compatible with the tree. Thus, in this case, Eq. (6) and Eq. (10) give the same result. K is also calculated correctly.

3. When all nodes have fossil calibrations, because in this case the conditional density $f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}} \mid \mathbf{t}_{\mathrm{c}}, t_1)$ does not need to be calculated.

In any other cases, the time prior needs to be examined explicitly. In the next section a couple of examples of calculation of the time prior under the correct and misspecified densities are given.

5 Some examples

Consider the four-species phylogeny of figure 1. The age of the root is $t_1 = 1$, and the age of node 2 is $t_2 = 0.2$. The age of node 3, t_3, is unknown. We want to construct a prior density on t_3 conditioned on t_1 and t_2 using the birth-death process: (1) first we find $f_{\mathrm{BD}}(t_2, t_3 \mid t_1)$; (2) then we find $f_{\mathrm{BD}}(t_2 \mid t_1)$; and (3) finally we find $f_{\mathrm{BD}}(t_3 \mid t_2, t_1) =$

$f_{\mathrm{BD}}(t_2,t_3 \mid t_1)/f_{\mathrm{BD}}(t_2 \mid t_1)$. The kernel density is $g(t_i \mid t_1) = 1/t_1$ (a uniform distribution between 0 and t_1) with distribution function $G(t_i \mid t_1) = t_i/t_1$. This is the limiting case when the parameters of the birth-death process are $\mu = \lambda, \rho = 0$ [2]. Using equations (3), (6), and (5), we get:

$$\begin{aligned} f_{\mathrm{BD}}(t_2,t_3 \mid t_1) &= 1/t_1^2, \\ f_{\mathrm{BD}}(t_2 \mid t_1) &= \int_0^{t_1} \frac{1}{t_1^2}\,\mathrm{d}t_3 = 1/t_1, \\ f_{\mathrm{BD}}(t_3 \mid t_2,t_1) &= \frac{f_{\mathrm{BD}}(t_2,t_3 \mid t_1)}{f_{\mathrm{BD}}(t_2 \mid t_1)} = 1/t_1. \end{aligned}$$

The same result can be obtained by noting that t_2 and t_3 are conditionally independent on t_1 (i.e. two random variables a and b are conditionally independent on c if $f(a,b \mid c) = f(a \mid c)f(b \mid c)$).

We now calculate the conditional density as currently implemented in MCMCTree, that is, by using Eq. 10:

$$f^*_{\mathrm{BD}}(t_2 \mid t_1) = \begin{cases} f_{\mathrm{BD}}(t_2 \mid t_1, t_2 > t_3) & \text{if } t_2 > t_3, \\ f_{\mathrm{BD}}(t_2 \mid t_1, t_2 < t_3) & \text{if } t_2 < t_3. \end{cases}$$

Applying Eq. (7) to calculate $f_{\mathrm{BD}}(t_2 \mid t_1, \pi)$ gives

$$f^*_{\mathrm{BD}}(t_2 \mid t_1) = \begin{cases} (s-2)!\,g(t_2 \mid t_1)G(t_2 \mid t_1) = 2t_2/t_1^2 & \text{if } t_2 > t_1, \\ (s-2)!\,g(t_2 \mid t_1)[1-G(t_2 \mid t_1)] = 2(t_1-t_2)/t_1^2 & \text{if } t_2 < t_3. \end{cases}$$

The misspecified prior of t_3 conditioned on t_2 and t_1 is then given by

$$f^*_{\mathrm{BD}}(t_3 \mid t_2,t_1) = \begin{cases} \frac{f^*_{\mathrm{BD}}(t_2,t_3|t_1)}{f_{\mathrm{BD}}(t_2|t_1,t_2<t_3)} = \frac{2}{t_1^2} \times \frac{t_1^2}{2(t_1-t_2)} = \frac{1}{t_1-t_2}, & \text{if } t_2 < t_3, \\ \frac{f^*_{\mathrm{BD}}(t_2,t_3|t_1)}{f_{\mathrm{BD}}(t_2|t_1,t_2>t_3)} = \frac{2}{t_1^2} \times \frac{t_1^2}{2t_2} = \frac{1}{t_2}, & \text{if } t_3 < t_2. \end{cases}$$

Using the fossil calibrations $t_1 = 1$ and $t_2 = 0.2$, we get

$$f^*_{\mathrm{BD}}(t_3 \mid t_2,t_1) = \begin{cases} f_{\mathrm{BD}}(t_3 \mid t_2 = 0.2, t_1 = 1, t_2 < t_3) = 1/(1-0.2) = 1.25 & \text{if } 0.2 < t_3 < 1, \\ f_{\mathrm{BD}}(t_3 \mid t_2 = 0.2, t_1 = 1, t_2 > t_3) = 1/0.2 = 5 & \text{if } 0 < t_3 < 0.2. \end{cases}$$

Note the density above is not a probability density as it does not integrate to one: $\int_0^{t_1} f_{\mathrm{BD}}(t_3 \mid t_2,t_1)\,\mathrm{d}t_3 = 2$ (this is because $K^* = 2$ is twice what it should be). Figure 1 shows the misspecified density $1/2 f^*_{\mathrm{BD}}(t_3 \mid t_2,t_1)$ and the correct density $f_{\mathrm{BD}}(t_3 \mid t_2,t_1)$. The shape of the misspecified density is confirmed numerically by MCMC sampling with MCMCTree. The shape of the misspecified density is not reasonable. Inadvertently, the user has specified an informative prior on t_3, with half of the prior probability mass on the narrow 0 to 0.2 interval, while a diffuse prior over the 0 to 1 interval was required.

Figure 3 shows an example where the resulting misspecified conditional prior is bimodal. Nodes 1–5 in the tree have point calibrations, while the age of node 6 is unknown. Using the uniform kernel density, it is easy to see that the correct density of

t_6 conditioned on the calibrated times is simply $1/t_1$. However, under the procedure of Eq. (10), the conditional density must be calculated over each one of the 5 labelled histories on the tree. The resulting misspecified density is thus a piecewise uniform distribution, with each segment of the distribution representing one labelled history, and with the resulting distribution having two modes (figure 3). Multimodal time priors like this may not be biologically realistic.

6 Integrating over histories: Calculating the joint density of $\mathbf{t}_c$

In small phylogenies (as in the four-species case), the marginal density $f_{\mathrm{BD}}(\mathbf{t}_c \mid t_1)$ can be obtained by solving the integral of Eq. (6) directly. For large phylogenies, the integral may be too cumbersome. Eq. (8) offers an alternative, by partitioning the integral as a sum over the labelled histories, with each integral having a known form (the joint density of a subset of order statistics, Eq. 7). However, for large phylogenies, the number of labelled histories may be too large to make computation of this sum practical. Here I discuss a post-order tree traversal algorithm to calculate the integral that may provide a way forward.

Before laying out the algorithm, it is useful to note the following. Consider the two daughter nodes of the root. These nodes are the last ancestors of two subtrees, which we call the left and right subtrees. For example, in the tree of figure 4, species a to f form the left subtree, while species g to j form the right subtree. The times on each subtree are conditionally independent on t_1, and thus the joint density of Eq. (3) can be written as

$$\begin{aligned} f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1) &= \frac{(s_R-1)!}{n_R}\prod g(t_i \mid t_1) \times \\ &\quad \frac{(s_L-1)!}{n_L}\prod g(t_j \mid t_1) \times \mathbb{I}_\tau, \end{aligned}$$

where s_R and s_L are the number of species on the right and left subtrees respectively ($s = s_R + s_L$), n_R and n_L are the number of labelled histories on the right and left subtrees, and the products are over the node ages on the right subtree (the t_i's), and over the left subtree (the t_j's). For example, for the tree of figure 4, $s_L = 6$, $s_R = 4$, $n_L = 8$, and $n_R = 1$. Conditional independence simplifies calculation of the integral of Eq. (6) as the non-calibrated node times in one subtree can be integrated out independently of the other subtree.

Now we can use a postorder algorithm to traverse the nodes of the ten-species phylogeny of figure 4 to integrate out the node ages without calibrations. If we start by visiting nodes on the left, the first node age to be integrated out is t_9. The partial integral is

$$\int_0^{t_7} g(t_9 \mid t_1)\,\mathrm{d}t_9 = G(t_7 \mid t_1).$$

The limits of integration are 0 and t_7 because $0 < t_9 < t_7$. We next visit node 8, and integrate t_8 out, giving

$$\int_0^{t_7} g(t_8 \mid t_1)\,\mathrm{d}t_8 = G(t_7 \mid t_1).$$

The algorithm now returns to node 7, and we integrate t_7 out:

$$\int_0^{t_5} g(t_7 \mid t_1) G^2(t_7 \mid t_1)\, \mathrm{d}t_7 = G^3(t_5 \mid t_1)/3.$$

The $G^2(t_7 \mid t_1)$ term inside the integral is the result of integrating the two daughter node ages, t_9 and t_8, in the previous steps. This integral is solved by recalling that $g(x) = G'(x)$. The algorithm now returns to node 5. The age of node 6 is not integrated out as it has a fossil calibration. Because node 6 has no daughter nodes we integrate t_5 directly:

$$\int_{t_6}^{t_1} \frac{1}{3} g(t_5) G^3(t_5 \mid t_1)\, \mathrm{d}t_5 = [G^4(t_1 \mid t_1) - G^4(t_6 \mid t_1)]/12.$$

Thus, because the left subtree is independent of the right subtree, and noting that $G^4(t_1 \mid t_1) = 1$, we get the marginal density of t_6 (one of the calibrated times) as

$$\begin{aligned} f_{\mathrm{BD}}(t_6 \mid t_1) &= \frac{(s_L - 1)!}{n_L} \times \frac{1}{12}[1 - G^4(t_6 \mid t_1)] \times g(t_6 \mid t_1), \\ &= 1.25 g(t_6 \mid t_1)[1 - G^4(t_6 \mid t_1)]. \end{aligned}$$

Now integrating out the non-calibrated node ages (t_2 and t_4) on the right subtree, we get the marginal density of t_3 (the other calibrated time) as

$$f_{\mathrm{BD}}(t_3 \mid t_1) = 6 G(t_3 \mid t_1) g(t_3 \mid t_1)[1 - G(t_3 \mid t_1)].$$

For example, if we set $t_1 = 1$, $g(t \mid t_1) = 1/t_1$, and $G(t \mid t_1) = t/t_1$, we get

$$f_{\mathrm{BD}}(t_3 \mid t_1) = 6 t_3 (1 - t_3), \tag{11}$$
$$F_{\mathrm{BD}}(t_3 \mid t_1) = 6(t_3^2/2 - t_3^3/3), \tag{12}$$
$$f_{\mathrm{BD}}(t_6 \mid t_1) = 1.25(1 - t_6^4), \tag{13}$$
$$F_{\mathrm{BD}}(t_6 \mid t_1) = 1.25(t_6 - t_6^5/5), \tag{14}$$

where F_{BD} are the appropriate cumulative distribution functions.

Figure 5 shows the marginal densities and distribution functions of Eqs. (11–14). To confirm the accuracy of the analytical calculations, we use MCMCTree to obtain samples from the joint distribution $f_{\mathrm{BD}}(\mathbf{t}_{\bar{1}} \mid t_1)$ (note this density is correctly calculated by MCMCTree). The sampled values of t_6 and t_3 can be summarised to obtain their distributions (histograms), or their sampled cumulative distributions. The sampled and analytical functions match almost perfectly (figure 5).

7 Discussion

The tree traversing algorithm laid out above can be implemented in a computer program by performing symbolic integration of the corresponding densities at the nodes of the tree. The symbolic solution to the integral can then be evaluated to perform MCMC sampling. My initial analysis suggest that all the possible integrals that can be generated have analytic solutions. However, the task of writing computer code to perform the symbolic integration may not be trivial. It may be worth exploring in detail the special case

of the uniform kernel density, $1/t_1$. A relatively simple general solution to the form of $f_{\mathrm{BD}}(\mathbf{t}_{\bar{\mathrm{c}}} \mid \mathbf{t}_{\mathrm{c}}, t_1)$ could perhaps be obtained under this kernel. Alternatively, Eq. (8), which has a known form, could be implemented in the program, but this would make MCMC feasible only for small trees, or for certain special types of large trees with few labelled histories. In the meantime the biologist interested in using MCMTree for Bayesian molecular clock dating must deal with the misspecified densities, unless the analysis can be performed under one of the three special cases laid out above.

The misspecification of the birth-death process in MCMCTree only affects the prior density of times without fossil calibrations, thus, the fossil calibration densities themselves are not affected. This means that in an analysis carried out using the misspecified density, the node ages will be adequately constrained by the fossil calibrations constructed by the user (however, note that truncation effects among fossil calibrations may affect the actual prior used, but this is entirely another issue [17]). Users of MCMCTree are advised to obtain MCMC samples from the prior (i.e. by running the program without sequence data), so that the prior can be examined. In most cases, the misspecified prior will be quite reasonable. Multimodal or other bizarre looking priors may be obtained with MCMCTree specially if several very precise calibrations are present throughout the tree. If the user considers the resulting priors to be biologically unrealistic, then several attempts at tweaking the calibrations and recalculating the prior by MCMC may provide a way forward. This advice should also be followed when estimating divergence times using any of the plethora of computer programs now available for Bayesian clock dating (e.g. [18, 19, 20, 21]): each program has its own idiosyncratic way of dealing with fossil calibrations, and unfortunately, it is not always possible to predict what the resulting priors will be.

Competing Interests

The author declares no competing interests.

Acknowledgements

Thanks to Ziheng Yang for extensive discussions, in particular, for his detailed explanation on how the time prior is calculated in his computer program, MCMCTree.

Funding

Part of this work was carried out while the author was a postdoc funded by Biotechnology and Biological Sciences Research Council (BBSRC, UK) grant BB/J009709/1 awarded to Ziheng Yang.

References

[1] Mario dos Reis, Philip CJ Donoghue, and Ziheng Yang. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17(2):71–80, 2016.

[2] Z. Yang and B. Rannala. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol*, 23(1):212–26, 2006.

[3] J. L. Thorne, H. Kishino, and I. S. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*, 15(12):1647–57, 1998.

[4] H. Kishino, J.L. Thorne, and W.J. Bruno. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution*, 18(3):352, 2001.

[5] A. J. Drummond, S. Y. Ho, M. J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88, 2006.

[6] J. Heled and A.J. Drummond. Calibrated trees priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.*, 61:138–149, 2012.

[7] Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):1586–91, 2007.

[8] R.W. Meredith, J.E. Janečka, J. Gatesy, O.A. Ryder, C.A. Fisher, E.C. Teeling, A. Goodbla, E. Eizirik, T.L.L. Simão, T. Stadler, et al. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science*, 334(6055):521–524, 2011.

[9] M. dos Reis, J. Inoue, M. Hasegawa, R. J. Asher, P. C. Donoghue, and Z. Yang. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc Biol Sci*, 279(1742):3491–500, 2012.

[10] Erich D Jarvis, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon YW Ho, Brant C Faircloth, Benoit Nabholz, Jason T Howard, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014.

[11] T. Stadler and Z. Yang. Dating phylogenies with sequentially sampled tips. *Syst Biol*, 62(5):674–88, 2013.

[12] F. Ronquist, S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn. A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. *Syst Biol*, 2012.

[13] R. D. Wilkinson, M. E. Steiper, C. Soligo, R. D. Martin, Z. Yang, and S. Tavare. Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Syst Biol*, 60(1):16–31, 2011.

[14] Tracy A Heath, John P Huelsenbeck, and Tanja Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966, 2014.

[15] Ziheng Yang and Bruce Rannala. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution*, 14(7):717–724, 1997.

[16] H.A. David and H.N. Nagaraja. *Order Statistics*. John Wiley & Sons, Hoboken, New Jersey, third edition edition, 2003.

[17] J. Inoue, P. C. Donoghue, and Z. Yang. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst Biol*, 59(1):74–89, 2010.

[18] N. Lartillot, T. Lepage, and S. Blanquart. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–8, 2009.

[19] T. A. Heath, M. T. Holder, and J. P. Huelsenbeck. A dirichlet process prior for estimating lineage-specific substitution rates. *Mol Biol Evol*, 29(3):939–55, 2012.

[20] F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Hohna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, 61(3):539–42, 2012.

[21] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

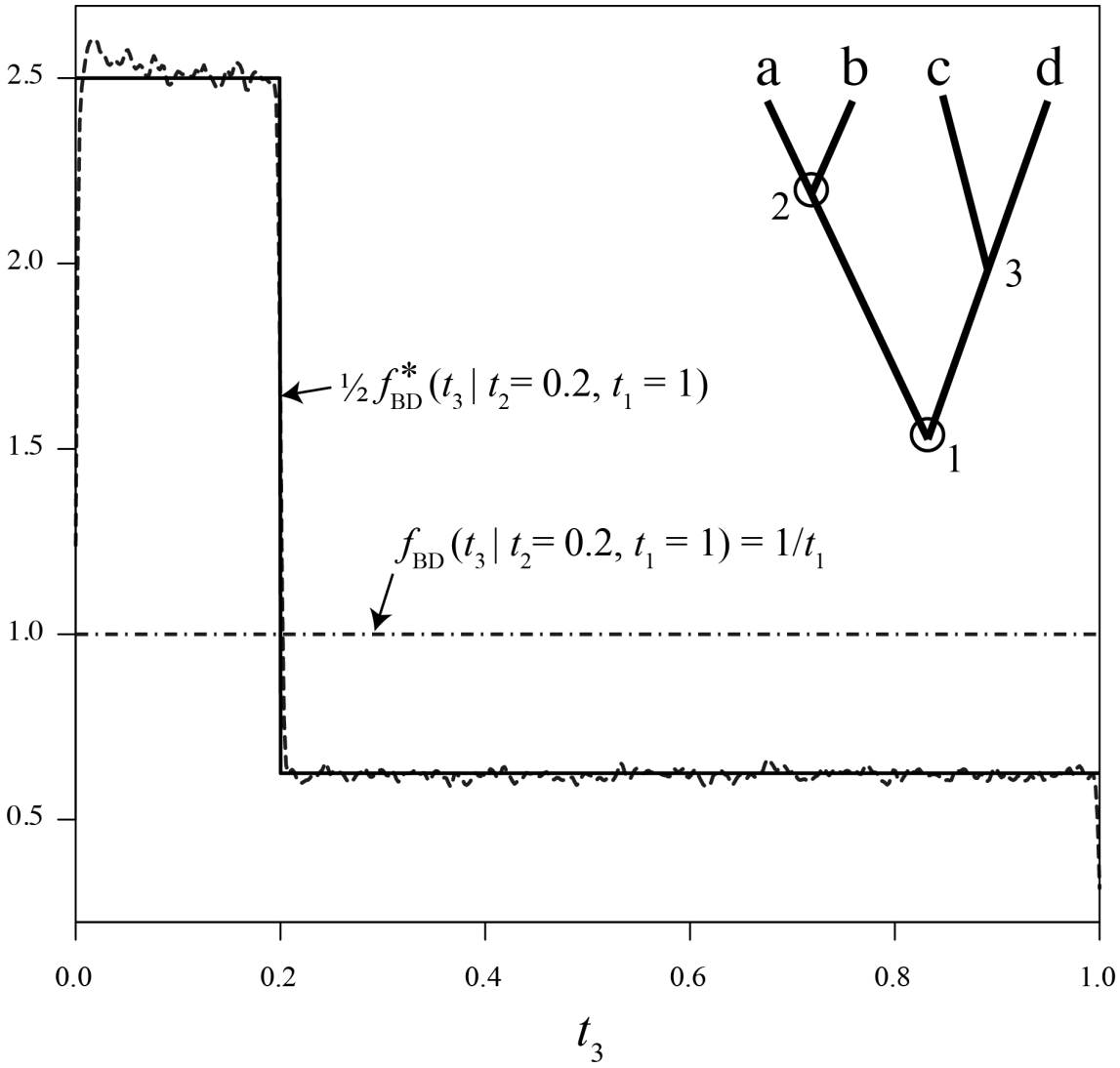

Figure 1: A misspecified birth-death prior with fossil calibrations. The inset tree has point fossil calibrations on the ages of nodes 1 and 2: $t_1 = 1$ and $t_2 = 0.2$ (white circles). The birth-death process with fossil calibrations is used to construct the prior of t_3 conditioned on the fossil ages. The correct conditional prior is $f_{\mathrm{BD}}(t_3 \mid t_2 = 0.2, t_1 = 1) = 1/t_1$ (dash-dotted line). The misspecified conditional prior, $f^*_{\mathrm{BD}}(t_3 \mid t_2 = 0.2, t_1 = 1)$, calculated under the procedure of Yang and Rannala [2] is a piecewise uniform distribution (solid line). The wiggly, dashed line shows the misspecified density sampled by MCMC using the computer program MCMCTree, which impements the mispecified prior. Because MCMCTree does not allow point calibrations, we use $t_2 \sim U(0.199, 0.201)$ and $t_1 \sim U(0.999, 1.001)$ as an approximation in the MCMC analysis. This example is analysed later in this article, where full details are given.

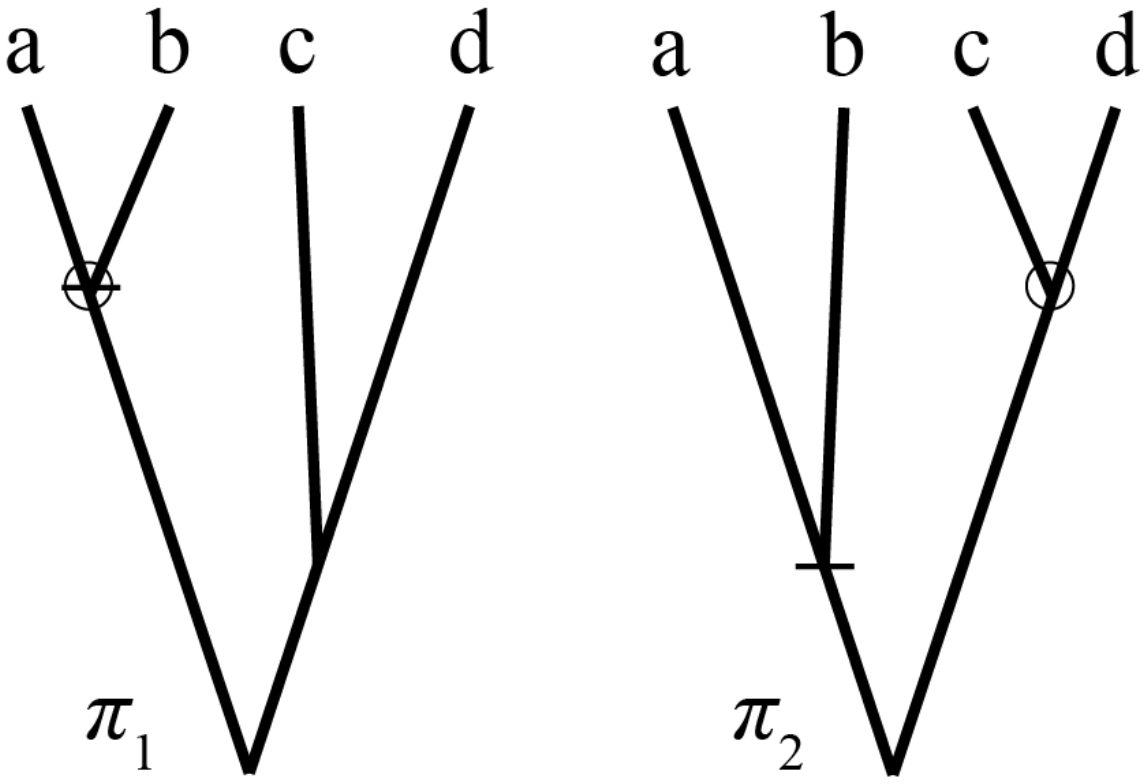

Figure 2: Two labelled histories for a four-species phylogeny. The labelled histories are denoted π_1 and π_2, and they share the same the tree topology τ =((a,b),(c,d)). The empty circle indicates the youngest node in the phylogeny, which has age $t_{(1)}$, while the small horizontal bar indicates the last common ancestor of a and b, which has age $t_{a,b}$.

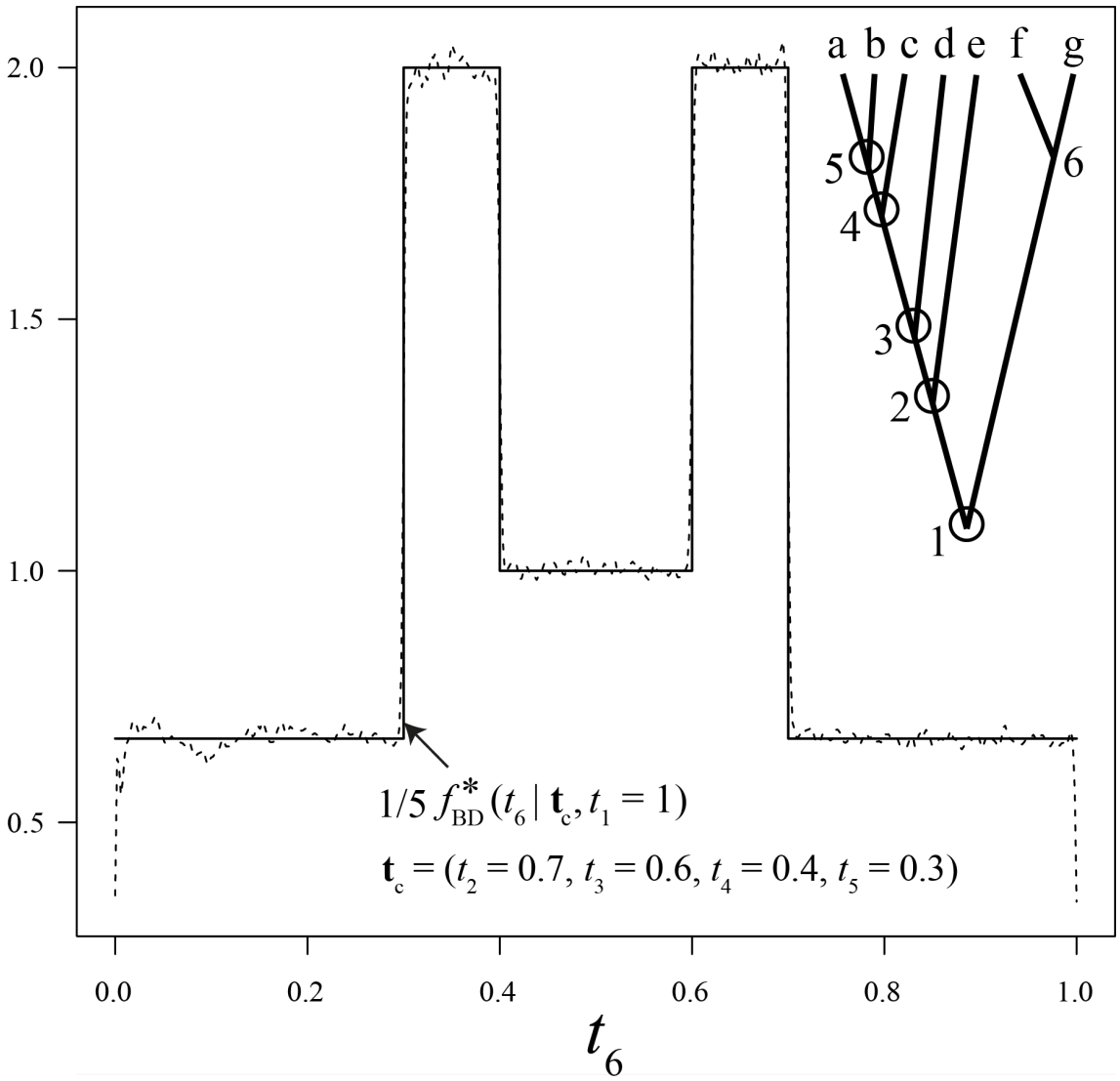

Figure 3: A misspecified, bimodal birth-death prior with fossil calibrations. The inset tree has point fossil calibrations for nodes 1 to 5: $t_1 = 1, t_2 = 0.7, t_3 = 0.6, t_4 = 0.4, t_5 = 0.3$ (white circles). The age of node 6 is unknown. The conditional prior of t_6 under the birth-death process is $f_{\mathrm{BD}}(t_6 \mid \mathbf{t}_c, t_1) = 1/t_1$ (not shown). The misspecified conditional prior, $f^*_{\mathrm{BD}}(t_6 \mid \mathbf{t}_c, t_1)$ (solid line), is a piecewise uniform distribution, where each segment corresponds to one of the 5 labelled histories compatible with the tree. The density has discontinuities located at the point fossil calibrations: 0.3, 0.4, 0.6 and 0.7. The mispecified normalising constant is $K^* = (s-2)! = 5!$. The correct constant is $K = (s-2)!/n_H = 5!/5$. Thus the mispecified density integrates to 5 (i.e. $K^*/K = 5$). The mispecified density is confirmed by MCMC sampling using MCMCTree (dashed line) using uniform calibrations: $t_i \sim U(c - 0.001, c + 0.001)$, where c is the calibration age for node i.

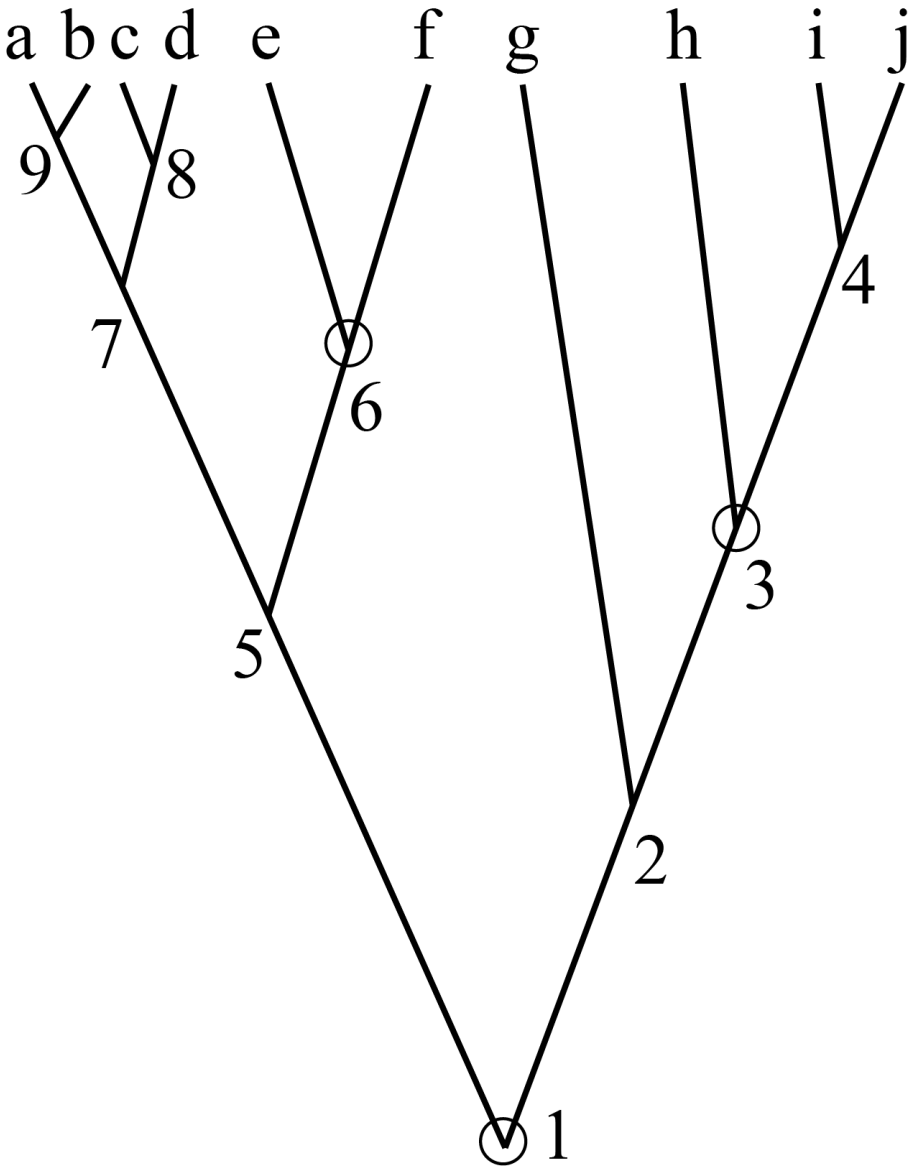

Figure 4: A ten-species tree with fossil calibrations. Nodes 1, 3 and 6 have fossil calibrations (white circles).

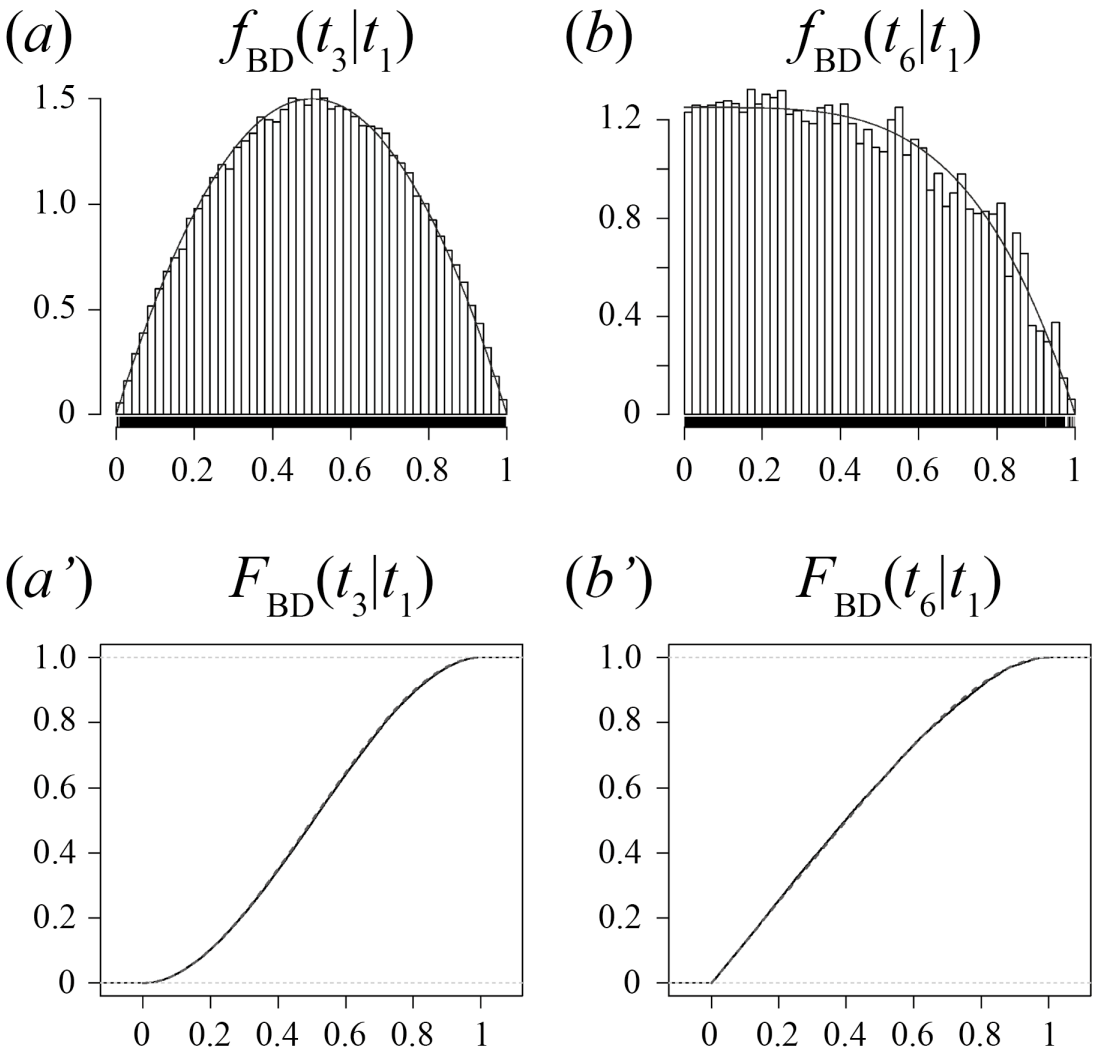

Figure 5: Marginal densities and distribution functions of two calibrated nodes in the ten-species phylogeny of figure 4. The marginal densities $f_{\mathrm{BD}}(t_3 \mid t_1)$ in (a), and $f_{\mathrm{BD}}(t_6 \mid t_1)$ in (b) are shown as solid lines. The corresponding sampled densities obtained with MCMCTree are shown as histograms. The cumulative distribution functions $F_{\mathrm{BD}}(t_3 \mid t_1)$ in (a'), and $F_{\mathrm{BD}}(t_6 \mid t_1)$ in (b') are shown as thick dashed lines. The sampled cumulative distributions obtained with MCMCTree are shown as solid lines (they overlap almost perfectly the analytical solutions). The analytic forms of the marginal densities and distribution functions are calculated according Eqs. (11–14).